9-2007

Forecasting the car penetration rate (CPR) in China: A nonparametric approach

Sainan JIN
Singapore Management University, snjin@smu.edu.sg

Liangjun SU
Singapore Management University, ljsu@smu.edu.sg

DOI: https://doi.org/10.1080/00036840600749631

Published in Applied Economics, Volume 39, Issue 17, September 2007, Pages 2189-2195
https://doi.org/10.1080/00036840600749631

Forecasting the car penetration rate (CPR) in China: a nonparametric approach

Sainan Jin* and Liangjun Su

Guanghua School of Management, Peking University, Beijing 100871, China

With strong economic growth, the auto industry has made great breakthroughs in recent years and has become a backbone industry in China, while cars play an increasingly important role, and are now the principal part of the auto industry. Both China's government and academic circles take strong interest in the prediction of CPR (i.e. car penetration rate or cars per thousand people), which will be the main guidance for the future industry policy. We summarize the existing problems in recent research and propose to use nonparametric methods to estimate the CPR and its elasticity with respect to GDP per capita (GDPPC). The results indicate that the nonparametric methods provide a much better fit than the conventional OLS method, and more importantly, it captures the nonlinearity of the elasticity of CPR with respect to GDPPC. Finally, we predict future CPR in China.

I. Introduction

With strong economic growth, rapid infrastructure development and the upcoming Beijing Olympics in the background, the auto industry has made great breakthroughs in recent years and has become a backbone industry in China. While companies and governments used to dominate the market of automobiles in the past, families and individuals have become the majority of the consumers nowadays. China produced about 1 million cars in 2002 and 2 million in 2003 (data source: China Automotive Industry Annual Book, 2005), constituting about 45.43% of the whole auto production. Cars play an increasingly important role and are now the principal part of the auto industry. Both China's government and academic circles take strong interest in the prediction of CPR (i.e. car penetration rate or cars per thousand people), which will be the main guidance for the future industry policy.

Despite the steady production growth, China's CPR is very low compared with the developed countries and most developing countries given its GDP per capita (abbreviated as GDPPC), or PPP GDP per capita (abbreviated as PPP GDPPC) (see Table 1 for the details). From basic economic theory, there exists a strong positive relationship between CPR and GDPPC (Yin and Gates, 2002). The higher the GDPPC, the higher the disposable income and the stronger the demand for cars will be. Many researchers predict CPR using this relation. But all of them assume that the elasticity of CPR with respect to GDPPC is constant, which is unrealistic. Further, in many applications, nonstationarity and spurious regression have not been taken account of while using time series or panel data.

*Corresponding author. E-mail: snjin28@yahoo.com

Table 1. CPR and GDPPC of selected countries (regions) 2001[a]

Country	APR[b]	CPR	GDPPC (USD)	PPP GDPPC (USD)
China	10.7	5.8	982	4600
South Korea	277.8	188.7	15 682	19 400
Japan	578	421.6	49 599	26 566
India	12.2	5.1	535	2540
Indonesia	26.5	14.7	1156	3000
Malaysia	263.2	212.8	5262	9000
Philippines	32.3	28.6	1319	4000
Thailand	106.4	43.1	3199	6600
USA	784	473	34 741	34 741
Canada	571.8	548.4	25 482	28 567
Germany	582.9	539.3	36 250	26 880
UK	543.5	484	24 993	25 552
France	584.4	484.8	33 755	26 340
Italy	641.2	576.1	23 529	25 957
Netherlands	461.8	408.7	34 684	28 467
Belgium	514.5	454.8	34 251	27 255
Spain	548.2	446	19 624	20 932
Sweden	497.9	451.8	36 601	25 348
Brazil	94.8	77.9	5111	7634
Australia	625.1	507	26 618	27 541

Notes: [a]Data source of cars: World Motor Vehicle Statistics 2003; data source of GDP: World Bank.
[b]APR is the abbreviation of automobile penetration rate.

We summarize the existing problems in recent research and point out that the elasticity of CPR with respect to GDPPC is not constant. Using panel unit root and panel cointegration tests, we find that both CPR and GDPPC are nonstationary and the two processes are not cointegrated, which implies that we could not use time series or panel data to study the relationship between the two processes, but the cross-sectional framework is not appropriate either, due to the varying CPR elasticity. We propose to use nonparametric methods to estimate CPR and its elasticity with respect to GDPPC. The results indicate that the nonparametric methods provide a much better fit than the conventional ordinary least square (OLS) method and more importantly, it captures the nonlinearity of the elasticity of CPR with respect to GDPPC. Finally, we predict future CPR in China.

The rest of the article is organized as follows. Section II performs panel unit root and panel cointegration tests to examine whether the CPR and GDPPC are nonstationary processes and whether they are cointegrated. Section III introduces nonparametric methods to estimate the model and the final section forecasts the CPR in China.

II. Existing Problems

A vast amount of literature has made the prediction of the CPR of China or other countries, but most of them are not open to the public. We summarize the existing literature based on published articles. There are three approaches to predicting the CPR, namely, the cross-sectional approach (Yang and Peng, 1994), the time series approach (e.g., Zheng and Wu, 1996); and most recently the panel data approach (e.g., Yin and Gates, 2002).

The cross-sectional approach is usually based upon a linear regression model and it gives a good estimation of the relationship between the CPR and GDPPC, but it ignores an important fact: the elasticity of CPR with respect to GDPPC is not constant over the range of GDPPC, thus the model is not appropriate for forecasting. The time series or panel data approach does not take account of the fact that both CPR and GDPPC are integrated processes and the two might not be cointegrated. Further, they also ignore the varying feature of the CPR elasticity.

Panel unit root and cointegration tests

Recently much attention has been given to panel unit root and panel cointegration research, which combines the information from time series as well as cross-sectional dimensions and provides useful econometric tools for empirical economic applications (Levin and Lim, 1992, two tests denoted as LL92a and LL92b; Levin and Lim, 1993, one test denoted as LL93) Im *et al.*, 1997, (TPS); Pedroni, 1999; McCoskey and Kao, 1998; Westerlund, 2005). For an overview of the literature, see Banerjee (1999) or more recently Westerlund (2005) and Su *et al.* (2005a, b).

We collect a panel data of CPR, GDPPC and PPP GDPPC of 12 developed countries (Australia, Belgium, Canada, France, Germany, Italy, Japan, Netherlands, Spain, Sweden, UK and USA) from 1970–2001 due to data availability. Before conducting panel unit root tests, we perform an ADF unit root test for each country[1] and find all time series processes exhibit a unit root at 5% significance level. Then we implement the panel unit root tests proposed by LL92a, LL92b, LL93 and IPS. The results are presented in Table 2.

From Table 2, all four tests show that both PPP GDPPC and CPR in logarithmic form are integrated of order one. Thus, it is necessary for us to

[1] We suspect there might be structural changes in some years and the rank-based unit root test by Cook (2004) gives similar results in the presence of structural breaks.

Table 2. Panel unit root tests results[a]

	Log(CPR)		Log(PPP GDPPC)	
Tests	Statistics	*p*-value	Statistics	*p*-value
LL92(a)	0.755	0.775	1.516	0.935
LL92(b)	−1.374	0.085	−0.123	0.451
LL93	3.697	0.999	3.579	0.999
IPS	5.108	1.000	4.650	1.000

Notes: [a]The null of the tests is that unit root exists. Under the null, the test statistic is asymptotically $N(0,1)$. The test results with log(GDPPC) are similar and will not be reported here.

Table 3. Panel cointegration tests results[a]

		Statistics	*p*-value
Pedroni (1999)	Panel ν	1.389	0.082
	Panel ρ	−0.935	0.175
	Panel t(np)[b]	−1.909	0.028
	Panel t(p)[c]	0.269	0.606
	Cluster ρ	0.416	0.661
	Cluster t(np)	−1.348	0.089
	Cluster t(p)	0.591	0.723
McCoskey and Kao (1998)	LM–FM	−1.256	0.893
	LM–DOLS	3.457	0.000
Westerlund (2005)	CUSUM–FM	−1.282	0.900
	CUSUM–DOLS	−1.167	0.878

Notes: [a]See McCoskey and Kao (1998); Pedroni (1999) and Westerlund (2005) for the definitions of the test statistics. The null of the seven test statistics from Pedroni (1999) is no cointegration, and under the null the test statistics are asymptotically normal. The null of the four test statistics from McCoskey and Kao (1998) and Westerlund (2005) is that cointegration exists, and under the null, the test statistics are asymptotically normal. The lag of DOLS test is set to be three. The test results for the relationship between log(CPR) and log(GDPPC) are similar and will not be reported here.
[b]np represents the nonparametric method.
[c]*p* represents the parametric method.

perform panel cointegration tests before regression analysis. We use seven test statistics constructed by Pedroni (1999), two test statistics constructed by McCoskey and Kao (1998) and one test statistic from Westerlund (2005). The results are shown in Table 3.

From Table 3, we can see that the test results are largely consistent with each other and they show that there are no cointegration relationships (at 5% significance level) between the two variables. We also employ the Johansen cointegration test with the optimum lags suggested by Bahmani-Oskooee and Brooks (2003) and find that at 5% significance level, cointegration does not exist in most countries (except Italy, Netherlands and Sweden) and at 1% level only the Netherlands supports the cointegration hypothesis. Therefore, we might encounter spurious regression if we use time series or panel data to analyse the relationship between CPR and the GDPPC.[2]

Constant elasticity?

We have emphasized that the assumption of constant elasticity of CPR with respect to GDPPC is not realistic. Figure 1 depicts the relationship between log(CPR) and log(PPP GDPPC) of some countries and regions from 1970–2001. We could see that the two variables do not exhibit a linear relationship in most countries. Thus, the classical linear model (including cross section, time series and panel data models) could not accurately describe the relationship between the two variables. This motivates us to consider nonparametric regression models in the next section.

III. Nonparametric Estimation

We consider a nonparametric regression model of the form:

$$y = m(x) + \varepsilon \tag{1}$$

where $y = \log(Y)$, $x = \log(X)$, with Y the CPR and X the GDPPC. We use two nonparametric methods to estimate $m(x)$: one is the Nadaraya–Watson (NW) kernel estimation, the other is local linear (LL) estimation. The former could be obtained by the following formula

$$\hat{m}(x) = \frac{\sum_{i=1}^{n} K_h(x - x_i)y_i}{\sum_{i=1}^{n} K_h(x - x_i)} \tag{2}$$

where n is the sample size and $K_h(u) = h^{-1}K(u/h)$, $K(\cdot)$ is a kernel function. The bandwidth $h = h(n) \to 0$ as $n \to \infty$. The latter approach sets $\tilde{m}(x) = \tilde{\beta}_0(x)$ and $(\tilde{\beta}_0(x), \tilde{\beta}_1(x))$ minimizes the following objective function:

$$\sum_{i=1}^{n} \{y_i - \beta_0 - \beta_1(x_i - x)\}^2 K_h(x - x_i) \tag{3}$$

[2] Leybourne and Newbold (2003) pointed out that spurious rejections, indicating the presence of cointegration, might occur in some cointegration tests with structural breaks. The finding of cointegration in the Netherlands by Johansen test might be induced by structural breaks.

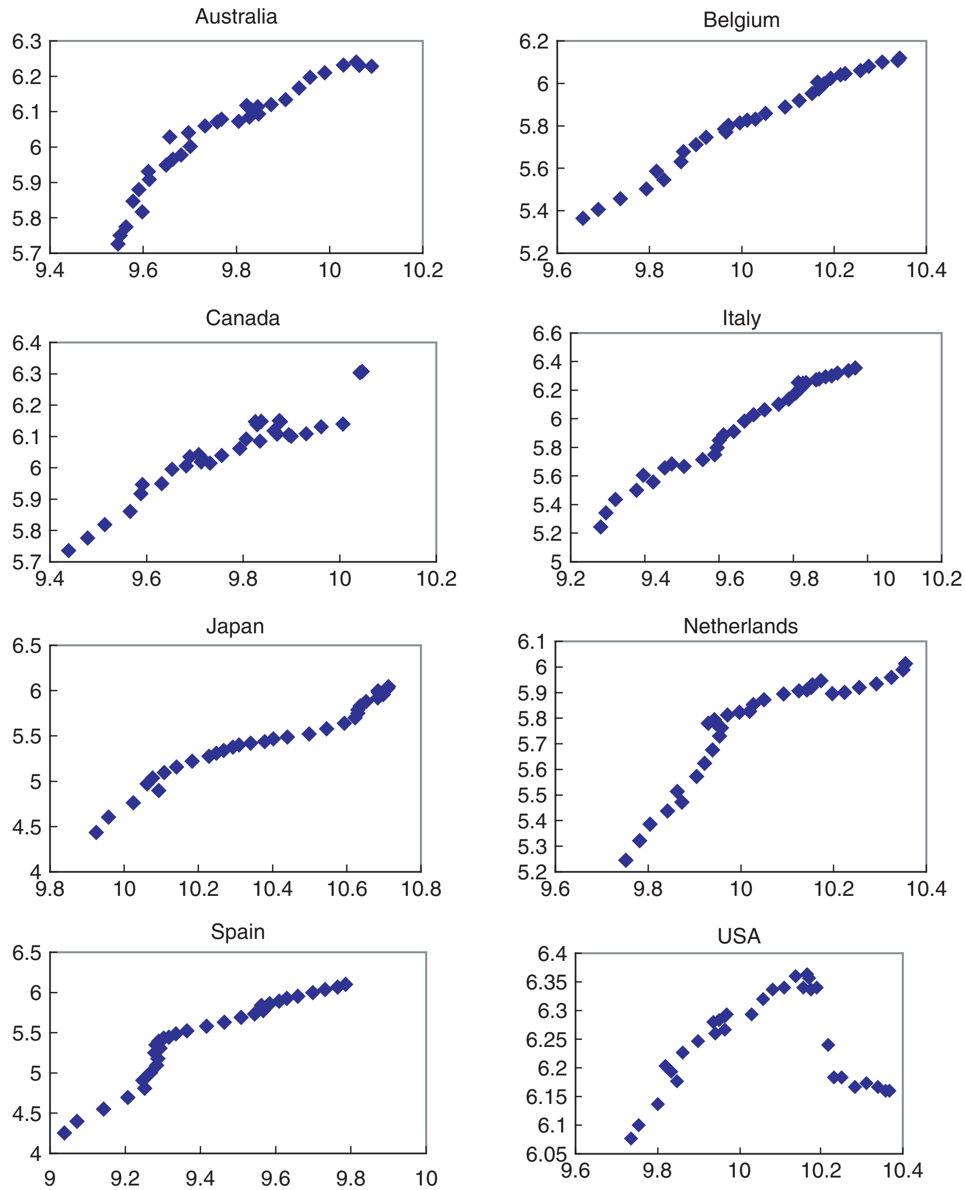

Fig. 1. Scatter plot of log(CPR) and log(PPP GDPPC)

Compared with the NW kernel estimation, the LL estimation has many advantages (Fan, 1992; Fan and Gijbels, 1996; Pagan and Ullah, 1999). It could correct the boundary bias automatically and at the same time give a consistent estimator $\tilde{\beta}_1(x)$ of the first derivative of the regression function, which we denote as $m'(x)$. This property is very important for empirical applications and is one of the main reasons that we employ the nonparametric methods. From (1) we see that $m'(x)$ represents the elasticity of CPR with respect to GDPPC.

We use cross-sectional data of 142 countries and regions in 2001 due to data availability. Table 4 presents some descriptive statistics. It shows that GDPPC varies a lot among countries, the maximum is about 441 times greater than the minimum; CPR has more variation, with the maximum 1137 times greater than the minimum.

We use standard normal kernel, and the bandwidth is chosen with the cross-validation approach: $h = c\hat{\sigma}_x n^{-1/5}$, where $\hat{\sigma}_x$ is the SE of x and $c \in \{0.1, 0.2, \ldots, 5\}$. We choose c to minimize the following function

$$\sum_{i=1}^{n} \{y_i - \hat{m}^{-i}(x_i)\}^2 w(x_i) \tag{4}$$

where $w(x_i) = 1\{|x_i| \le 2\hat{\sigma}_x\}$ is the weight function and $\hat{m}^{-i}(x_i)$ is the leave-one-out estimator of $m(x_i)$ by using $h = c\widehat{\sigma}_x n^{-1/5}$.

Table 4. Descriptive statistics of data

	Mean	SE	Min	Median	Max	Skewness	Kurtosis
CPR	139	179	0.549	44.8	625	1.24	0.207
GDPPC	7648	11 137	101	2332	44 514	1.77	2.044

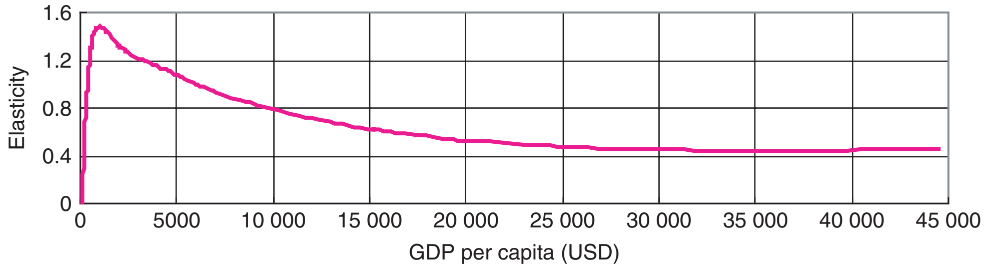

Fig. 2. CPR elasticity

Compared with the OLS estimation, the nonparametric methods provide a much better goodness-of-fit, the R^2 rises from 0.79 to 0.95 (for the NW case) or 0.96 (for the LL case). The regression results are very similar for these two nonparametric methods. We also depict the scatter plot of CPR elasticity against GDPPC in Fig. 2. The elasticity is not constant at all. When the income is low, the elasticity is relatively low too; when GDPPC rises above 250 USD, the elasticity increases dramatically; when GDPPC reaches about 1000 USD, the CPR elasticity reaches the summit. Then it has a downward trend until GDPPC arrives at about 30 000 USD, where the elasticity remains at the level of 0.45. The OLS estimation of CPR elasticity appears to be relatively coarse and actually it only reflects the average level of the CPR elasticity of all countries. But every country or region has different levels of GDPPC, the CPR elasticity will surely be very different among countries and regions. Therefore it is important for us to take account of the fact that the CPR elasticity is not constant over the range of GDPPC when we want to forecast CPR.

IV. Forecasting the CPR in China

The auto market in China has been influenced greatly by the government policy, so it is relatively difficult to predict the actual CPR in China, whereas it is feasible to forecast the 'reasonable' CPR given a certain GDPPC level. The prediction of this 'reasonable' level of CPR is valuable for both policy evaluations and economic research applications.

First we calculate the 'reasonable' CPR in China from 1990 to 2003 based on international comparisons and compare it with the actual CPR data. It is well known that GDPPC calculated using nominal exchange rate normally underestimates the economy level of developing countries, whereas GDPPC based on PPP normally overestimates it. The purpose of this article is to point out that the auto market of China is lagging behind and has enormous potential, thus, we employ the more conservative approach, i.e. using GDPPC based on nominal exchange rate here. Table 5 gives the comparison results. It also presents the estimation results with OLS estimation so that we could have a better sense about the obvious advantage of nonparametric methods.

We make some remarks in turn.

(1) The sedan market is lagging far behind from its 'reasonable' level. For example, in 1990, the actual CPR in China was 1.4, less than 30% of the predicted 'reasonable' level; in 2001, the actual CPR is 5.8, only half of the predicted 'reasonable' level.
(2) The OLS prediction will underestimate the 'reasonable' CPR at the preliminary stage of the sedan market development and tend to overestimate it when the sedan market is mature. This is misleading. The GDPPC growth rate is about 8.6% from 1990 to 2003 and the predicted CPR elasticity with OLS regression is around 1.07, so the predicted 'reasonable' CPR growth rate is around 8.6%. However, from Fig. 2, we can tell that corresponding to this period of GDPPC, the CPR elasticity rises from 1.07 to 1.47 rapidly and remains above 1.4 for

Table 5. Forecasting of 'reasonable' CPR (1990–2003)

Year	Real CPR	GDPPC (USD)	OLS	NW	LL
1990	1.4	391	3.4	4.6	4.0
1991	1.6	421	3.7	4.9	4.3
1992	2.0	475	4.1	5.3	4.9
1993	2.3	533	4.6	5.9	5.6
1994	2.6	593	5.2	6.7	6.5
1995	2.8	649	5.6	7.4	7.3
1996	3.1	704	6.1	8.2	8.2
1997	3.9	758	6.6	9.1	9.1
1998	4.1	809	7.0	10.0	10.0
1999	4.7	859	7.5	11.0	10.9
2000	4.6	920	8.0	12.2	12.1
2001	5.8	982	8.5	13.6	13.3
2002		1054	9.2	15.2	14.8
2003		1142	9.9	17.3	16.7
Average annual growth rate[a]	13.6%	8.6%	8.6%	10.7%	11.5%

Notes: [a]The average annual growth rate of the real CPR is calculated for 1990–2001, and the growth rate of others is calculated for 1990–2003.

another 8 years. The CPR elasticity based on the OLS estimation remains constant and it could hardly capture the rapid increase in the CPR elasticity due to the economic growth during the developing stage of the sedan market; neither could it capture the downward trend of CPR elasticity when the sedan market is mature.

(3) The sedan market in China has a huge potential. We could forecast the CPR in the next middle-to-long run. It is expected that starting with the year 2005, the GDPPC will be quadrupling for the first time in 2020 and for the second time in 2050 according to the development planning of China's state government. Therefore we assume the growth rate of GDPPC is 7% during 2005–2020 and 5% during 2020–2050. The prediction results are displayed in Table 6.

As seen in Table 6, the prediction results based on OLS estimation are quite different from the results based on nonparametric estimation. We will focus on the nonparametric prediction results. If we assume on average, a household is composed of 3.3 people, then about every 10 households will own a car in 2010. This 'reasonable' prediction might be difficult to accomplish due to the backward development of the sedan market. In 2020, every one-fifth to one-fourth households will own a car and in 2050, almost every household will own a car. At that time cars will

Table 6. Long-run forecasting of 'reasonable' CPR

Year	Forecasting based on GDPPC GDPPC(USD)[a]	OLS	NW	LL
2005	1308	11.4	21.3	20.3
2010	1835	16.0	34.1	32.8
2015	2573	22.4	50.2	50.9
2020	3609	31.4	70.5	76.6
2030	5878	51.1	113.6	128.7
2040	9575	83.3	184.2	196.0
2050	15 597	135.6	269.4	273.7

Note: [a]GDPPC in 2001 dollars.

become normal goods just like mobile phones in today's China. The earlier prediction also conforms with the experience of some developed countries: when GDPPC reaches 16 000 to 18 000 USD, the cars per 1000 people will be around 250–400. Therefore, our long-run 'reasonable' prediction is possible to realize in the future.

Acknowledgement

The authors thank the NSFC for research support under grants 70501001 and 70601001.

References

Bahmani-Oskooee, M. and Brooks, T. J. (2003) A new criteria for selecting the optimum lags in Johansen's cointegration technique, *Applied Economics*, **35**, 875–80.

Banerjee, A. (1999) Panel data unit roots and cointegration: an overview, *Oxford Bulletin of Economics and Statistics*, **61**, 608–28.

Cook, S. (2004) Rank-based unit root testing in the presence of structural change under the null: simulation results and an application to US inflation, *Applied Economics*, **37**, 607–17.

Fan, J. (1992) Design-adaptive nonparametric regression, *Journal of the American Statistical Association*, **87**, 998–1004.

Fan, J. and Gijbels, I. (1996) *Local Polynomial Modelling and Its Applications*, Chapmman and Hall, London.

Im, K. S., Pesaran, M. H. and Smith, R. (1997) Testing for unit roots in heterogeneous panels, *mimeo*, Department of Applied Economics, University of Cambridge.

Levin, A. and Lin, C. F. (1992) Unit root tests in panel data: asymptotic and finite sample properties, Discussion Paper, Department of Economics, UCSD.

Levin, A. and Lin, C. F. (1993) Unit root tests in panel data: new results, Discussion Paper, Department of Economics, UCSD.

Leybourne, S. J. and Newbold, P. (2003) Spurious rejections by cointegration tests induced by structural breaks, *Applied Economics*, **35**, 1117–21.

McCoskey, S. and Kao, C. (1998) A residual-based test for the null of cointegration in panel data, *Econometric Reviews*, **17**, 157–66.

Pagan, A. and Ullah, A. (1999) *Nonparametric Econometrics*, Cambridge University Press, Cambridge.

Pedroni, P. (1999) Critical values for cointegration tests in heterogeneous panels with multiple regressions, *Oxford Bulletin of Economics and Statistics*, **61**, 653–70.

Su, L., He, Y. and Jin, S. (2005a) Does temporary income really affect consumption?—evidence from a rural panel study in China, *Management World*, **7**, 26–30.

Su, L., He, Y. and Jin, S. (2005b) A panel cointegration study of the relationship between income and consumption in China, *World Economy*, **333**, 65–72.

Westerlund, J. (2005) A panel CUSUM test of the null of cointegration, *Oxford Bulletin of Economics and Statistics*, **67**, 231–60.

Yang, X. and Peng, Y. (1994) Analysis of China auto industry, *Journal of Quantitative and Technical Economics*, **7**, 17–39.

Yin, J. Z. and Gates, D. F. (2002) Automobile and fuel industries, in *The Globalization of the Chinese Economy* S.-J. Wei, G. J. Wen and H. Zhou (Eds), Edward Elgar, Cheltenham, pp. 82–99.

Zheng, D. and Wu, Y. (1996) Prediction of CPR with exponential smoothing, *Industrial Technology & Economy*, **15**, 81–83.